

Supplementary Materials

Domain-Specific Fine-Tuning of Large Language Models and Intelligent Question-Answering System for Industrial Catalysis

Shican Wu^{1,2,†}, Xin Chang^{1,2,†}, Xiao Ma^{1,2}, Xiaoyun Lin^{1,2}, Ran Zhao^{1,2} and Zhi-Jian Zhao^{1,2,*}¹ Key Laboratory of Green Synthesis and Transformations, Ministry of Education, School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China² International Joint Laboratory of Low-Carbon Chemical Engineering, Ministry of Education, Tianjin 300350, China

* Correspondence: zjzhao@tju.edu.cn

† These authors contributed equally to this work.

How To Cite: Wu, S.; Chang, X.; Ma, X.; et al. Domain-Specific Fine-tuning of Large Language Models and Intelligent Question-Answering System for Industrial Catalysis. *Smart Chemical Engineering* 2026, 2(1), 2. <https://doi.org/10.53941/sce.2026.100002>

1. Domain-Specific General Pre-training Corpus Screening and Integration

For continuous pre-training (CPT) corpus construction, this research first performed cleaning processing on English documents from the redpajama-refine dataset as the general pre-training foundation. To maintain model domain generalization performance, the screening process prioritized extraction of content related to the industrial catalysis field. As shown in Figure S1a, after systematic data cleaning, propane dehydrogenation literature related to industrial catalysis yielded a total of 9.13 M tokens, with approximately 215,000 entities extracted through entity extraction technology. These preliminarily extracted entities underwent large language model annotation, retaining 10,025 higher-frequency items, and finally 3768 high-quality domain-specific terms were determined through manual annotation, forming the core of the domain-specific vocabulary. These carefully screened professional terms were subsequently applied to relevance evaluation and screening of general corpora, ensuring training data possessed sufficient domain specificity.

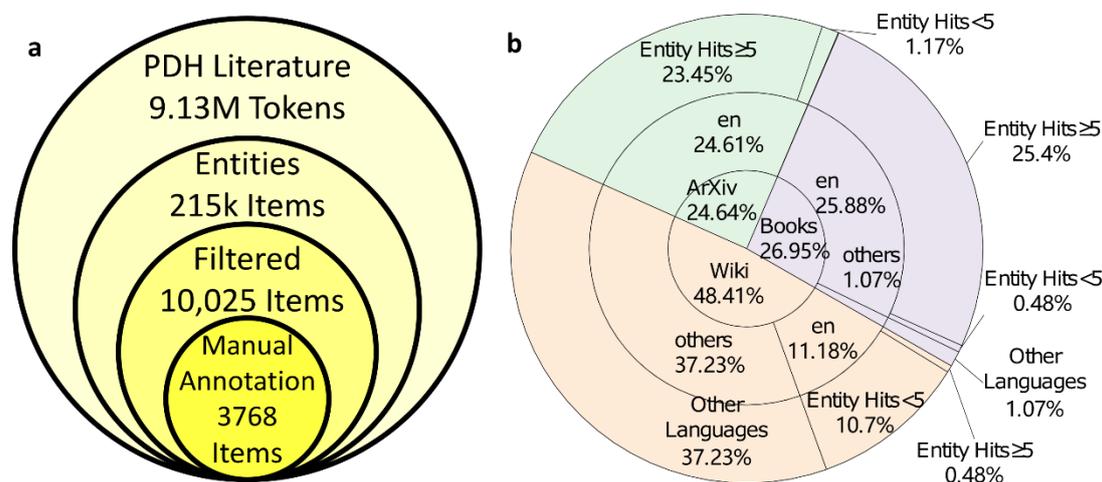


Figure S1. (a) Domain-specific vocabulary construction process (b) Sunburst chart showing proportion of redpajama-refine dataset after domain-specific vocabulary filtering, from inner to outer rings: (i) dataset subsets; (ii) language distribution; (iii) domain entity match distribution.

As shown in Figure S1b, distribution analysis of the redpajama-refine dataset reveals that, according to entity match statistics, content matching 5 professional entities accounts for the highest proportion, at 23.45% and 25.4% respectively. From content type perspective, Wiki-type materials account for 48.41% of the dataset, ArXiv-type academic content accounts for 24.64%, and Book-type content accounts for 26.95%. Regarding language



distribution, English content totals 61.67% (comprising 24.61%, 25.88%, and 11.18% in three parts). In the actual sampling process, this research primarily sampled from document collections matching more than 5 entities across the three major categories of ArXiv, Book, and Wiki, to improve corpus domain relevance.

In addition to the aforementioned English resources, CPT general corpus also integrated Chinese Wikipedia data and high-quality Chinese resources such as CSL (Chinese Scientific Literature), constructing a multilingual pre-training foundation. To further enhance the domain specificity of training data, the research established a screening mechanism based on professional vocabulary: Wiki and ArXiv content required at least 5 domain-specific terms, while book-type resources were held to higher standards, requiring more than 49 relevant professional terms per entry. This domain vocabulary-based screening strategy effectively improved the relevance between general corpora and the domain, providing more precise learning materials for subsequent model training. Through the above systematic screening and integration process, a high-quality general CPT corpus containing 0.753B tokens was ultimately constructed.

2. Base Model Selection Benchmark Comparison

Figure S2 presents the comprehensive benchmark comparison of Yi-1.5-6B against contemporaneous open-source models of comparable scale (6B-8B parameters), as reported by the Yi-1.5 model card (<https://huggingface.co/01-ai/Yi-1.5-6B> (accessed on 12 May 2024)). Yi-1.5-6B achieves the most balanced performance profile among all compared models, with leading scores in CMMLU (70.8), AGIEval (56.0), MBPP+ (46.9), and Math (28.42), while maintaining competitive performance on English benchmarks (MMLU: 63.5, BBH: 59.0). Notably, Yi-1.5-6B significantly outperforms other same-scale models in Chinese language understanding (CMMLU: 70.8 vs. Gemma-7B: 48.4, Mistral-7B: 44.6, Llama-3-8B: 50.9), which is critical for our bilingual training strategy.

Model	MMLU	CMMLU	BBH	AGIEval	HumanEval (+)	MBPP (+)	GSM8k	Math
Yi-1.5-9B	69.5	74.8	72.4	62.7	41.4(34.1)	61.1(53.6)	73.7	32.6
Yi-1.5-6B	63.5	70.8	59.0	56.0	36.5(28.7)	56.8(46.9)	62.2	28.42
Gemma-7B	64.3	48.4	61.0	46.0	33.5(28.0)	45.8(32.8)	55.7	24.8
Qwen1.5-7B	61.0	73.4	13.3	61.6	36.0(31.1)	46.1(37.6)	70.1	20.3
Mistral-7B	62.5	44.6	58.2	42.4	29.3(22.6)	50.2(32.1)	47.5	15.5
Mixtral-8x7B-v0.1	70.6	53.0	70.5	49.5	40.2(31.1)	60.7(31.1)	65.7	28.4
Llama-3-8B	66.6	50.9	64.4	44.7	34.7(31.7)	48.0(44.9)	54.7	21.16

Global open-source LLM English/Chinese benchmarks

Figure S2. Benchmark comparison of Yi-1.5-6B and Yi-1.5-9B base models against contemporaneous open-source LLMs on English and Chinese benchmarks (as of May 12, 2024). Data source: <https://huggingface.co/01-ai/Yi-1.5-6B> (accessed on 12 May 2024).

3. Detailed Results of General Capability Evaluation

This research constructed a comprehensive general capability test scheme based on the OpenCompass evaluation framework, conducting a comprehensive evaluation of the PeiYang Micro-Emergence model's general capabilities and comparing with Yi-1.5-6B and Yi-1.5-6B-Chat. The evaluation covered multiple dimensions including language understanding, common sense memory, logical reasoning, and mathematical reasoning, comprising 39 standard test sets in total, forming a multidimensional characterization of model capabilities.

From the capability dimension perspective, the domain fine-tuned PeiYang Micro-Emergence model performed outstandingly in language understanding, particularly in text comprehension tasks (WiC, WSC) where it was comparable to Yi-1.5-6B-Chat, achieving 50.47 and 50.96 points respectively. In logical reasoning, the PeiYang Micro-Emergence model performed excellently, achieving 91.00 points on the COPA test set, exceeding Yi-1.5-6B-Chat's 88.00 points, demonstrating good causal reasoning capabilities. This result is related to training

reinforcement of causal reasoning understanding in the industrial catalysis field. We note that the improvement in causal reasoning represents a secondary outcome of domain training rather than the primary contribution, which is enabling the model to understand professional catalysis terminology and knowledge.

In terms of common-sense memory capability, the PeiYang Micro-Emergence model shows a gap compared to Yi-1.5-6B-Chat. On knowledge-intensive test sets such as MMLU and C-Eval, the PeiYang Micro-Emergence model scored 35.20 and 45.03 points respectively, while Yi-1.5-6B-Chat achieved 56.75 and 60.72 points. This difference reflects the model's attention shifting toward the catalysis field during domain training, with relatively reduced weight on general knowledge. Similarly, in mathematical reasoning tasks, there are also notable differences between the PeiYang Micro-Emergence model and Yi-1.5-6B-Chat. On GSM8K and MATH datasets, the PeiYang Micro-Emergence model scored 37.45 and 3.14 points respectively, while Yi-1.5-6B-Chat achieved 76.04 and 40.86 points. This difference stems from the domain training process where industrial catalysis corpora contain fewer structured mathematical reasoning tasks, causing relative weakening of the model in such capabilities.

Through comparison, the original Yi-1.5-6B base model generally performed poorly across all tests, with most scores below 20 points. This does not indicate that the base model lacks relevant knowledge representations, but rather that without SFT training, the base model has not acquired instruction-following capabilities and cannot output answers in the format required by the evaluation system, thus being judged as scoring zero. This also indirectly demonstrates the necessity of the CPT-SFT joint training strategy adopted in this research, which effectively enhanced the model's instruction understanding and execution capabilities.

Overall, the PeiYang Micro-Emergence model exhibits a general capability spectrum similar to Yi-1.5-6B-Chat, but with specific tilts in capability distribution. This tilt reflects, on one hand, the characteristics of industrial catalysis field training corpora, such as emphasis on concept explanation and causal analysis; on the other hand, it also demonstrates the effectiveness of the 1:1 domain-specific to general corpus ratio strategy in balancing specialization and generality. We further note that the mathematical reasoning decline (GSM8K: 37.45 vs. 76.04; MATH: 3.14 vs. 40.86) represents a known trade-off in domain-specific fine-tuning, as industrial catalysis corpora contain fewer structured mathematical reasoning tasks. In practical deployment, mathematical calculations can be performed through tool-calling capabilities, which were not enabled during evaluation. Moreover, this decline partly reflects a fundamental limitation of current LLM architectures: numbers are processed as token sequences rather than numerical values (e.g., "0.11" and "0.9" have tokenized representations unrelated to their magnitudes), making mathematical reasoning inherently challenging regardless of fine-tuning approach. Complete evaluation results are shown in Table S1 below:

Table S1. Comparative evaluation results on general benchmark datasets.

Dataset	Capability Dimension	PeiYang Micro-Emergence	Yi-1.5-6B	Yi-1.5-6B-Chat
ARC	Common Sense Memory	80.41	42.22	79.28
C3	Language Understanding	82.03	20.82	74.74
CB	Language Understanding	66.07	42.86	62.50
COPA	Logical Reasoning	91.00	48.00	88.00
MultiRC	Language Understanding	68.01	19.41	76.53
ReCoRD	Language Understanding	31.25	0.34	0.05
WSC	Language Understanding	50.96	4.81	51.92
WiC	Language Understanding	50.47	4.70	55.49
AFQMC	Language Understanding	70.51	8.55	62.44
AGIEval	Comprehensive Capability	47.92	15.86	47.44
BUSTM	Language Understanding	72.31	18.52	48.84
C-Eval	Common Sense Memory	45.03	6.17	60.72
CHID	Language Understanding	64.38	13.76	61.28
CLUEWSC	Language Understanding	81.11	44.25	76.09
CMB	Language Understanding	17.86	3.40	13.22
CMMLU	Common Sense Memory	44.31	10.76	66.71
CMNLI	Language Understanding	65.51	12.22	32.36
CSL	Language Understanding	49.69	23.75	49.17
EPRSTMT	Language Understanding	87.85	33.97	82.10
GSM8K	Mathematical Reasoning	37.45	0.30	76.04
HellaSwag	Language Understanding	53.10	28.26	69.77
LAMBADA	Language Understanding	38.50	0.85	41.37
LCSTS	Creative Generation	17.16	0.93	9.56
MATH	Mathematical Reasoning	3.14	10.30	40.86
MMLU	Common Sense Memory	35.20	1.53	56.75
NQ	Common Sense Memory	14.35	16.23	17.48
OCNLI	Language Understanding	60.00	16.44	36.44
OCNLI-FC	Language Understanding	66.79	24.97	40.52
HumanEval	Common Sense Memory	31.71	38.41	62.20
OpenBookQA	Common Sense Memory	77.90	41.20	74.50
PIQA	Common Sense Memory	74.16	36.78	66.59
RACE	Language Understanding	80.95	26.77	70.52

Table S1. Cont.

Dataset	Capability Dimension	PeiYang Micro-Emergence	Yi-1.5-6B	Yi-1.5-6B-Chat
SimpleQA	Logical Reasoning	69.75	42.78	38.89
StoryCloze	Language Understanding	94.07	57.99	91.02
StrategyQA	Logical Reasoning	59.08	65.59	63.19
SummEdits	Creative Generation	53.89	1.61	56.63
TNEWS	Language Understanding	52.88	11.05	44.41
TriviaQA	Common Sense Memory	35.34	26.37	41.78
WinoGrande	Language Understanding	59.59	28.10	64.56

The capability radar chart of evaluation results is shown in Figure S3, intuitively displaying the performance differences of the three models across various capability dimensions:

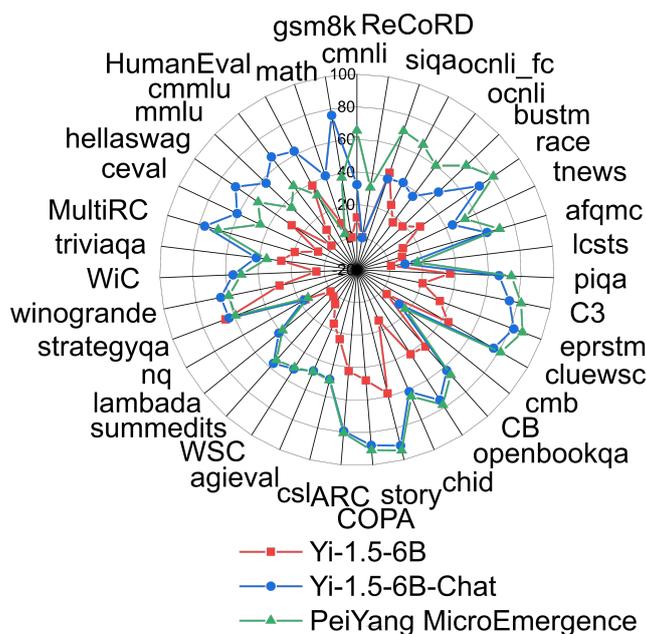


Figure S3. Radar chart of evaluation results on general benchmark datasets, with green line representing PeiYang Micro-Emergence model, red line representing fine-tuned base Yi-1.5-6B model, and blue line representing baseline Yi-1.5-6B-Chat model.

4. Theoretical Foundation of Matryoshka Representation Learning and MRL-based Embedding Model Fine-Tuning Method

Matryoshka Representation Learning (MRL) is a multi-granularity representation learning method, with its core idea being to encode multi-granularity representations in a single embedding vector, enabling models to flexibly adjust according to downstream task complexity and computational resources. MRL achieves the inclusion of multiple effective representations of different dimensions within the same high-dimensional vector through nested structure representation vectors.

In standard representation learning, for a set of representation vectors $M \subset \mathbb{N}$ with dimension d , MRL learns a representation vector $z \in \mathbb{R}^d$ such that for any $m \in M$, the first m dimensions of the vector $z_{1:m} \in \mathbb{R}^m$ can serve as an effective representation of the input data point x . MRL implements the mapping $z = F(x; \theta_F)$ through a neural network $F(\cdot; \theta_F): X \rightarrow \mathbb{R}^d$. The multi-granularity property is captured through the set M which $|M| \leq \lfloor \log(d) \rfloor$. In practice, MRL is implemented through multi-scale representation learning problems. Taking multi-classification tasks as an example, the MRL optimization objective is:

$$\min_{\{W^{(m)}\}_{m \in M}, \theta_F} \frac{1}{N} \sum_{i \in [N]} \sum_{m \in M} c_m \cdot L(W^{(m)} \cdot F(x_i; \theta_F)_{1:m}; y_i) \quad (S1)$$

where the loss function $L: \mathbb{R}^L \times [L] \rightarrow \mathbb{R}^+$ is the multi-class softmax cross-entropy loss, and the weights $c_m \geq 0$ represent the relative importance of each nested dimension.

The advantages of MRL include: (1) encoding multi-granularity information in a single vector, avoiding the need to maintain multiple models of different dimensions; (2) low-dimensional representation capability comparable to, or even better than, independently trained low-dimensional models; (3) obtaining multi-granularity

representation capability without additional inference cost; (4) strong adaptability, allowing dynamic adjustment of representation dimensions according to task requirements and computational resources.

When selecting the base embedding model, systematic evaluation was conducted on multiple pre-trained models, including bge-m3, bge-large-zh-v1.5, and bge-base-zh-v1.5. Comprehensively considering vector quality, inference speed, and resource requirements, bge-m3 was ultimately selected as the base model for fine-tuning. The main basis for selecting bge-m3 was its excellent multilingual capability, powerful semantic representation capability, and good transfer effects, which are particularly important for the bilingual research environment in the industrial catalysis field.

This research adopted a fine-tuning method based on Matryoshka Representation Learning (MRL), conducting secondary development on the percent4/embedding_model_exp repository to build training and evaluation code, thereby supporting inference and evaluation with the sentence_transformer library. The MRL-based fine-tuning method contains the following core steps: first, loss functions are separately defined for representation vectors of different dimensions to achieve multi-granularity representation learning. For training hyperparameters, batch size was set to 64, training epochs to 3, learning rate to 2×10^{-5} with linear decay strategy, optimizer selected as AdamW with weight decay 1×10^{-2} , and maximum sequence length to 1024. For contrastive learning configuration, temperature coefficient was set to 0.07, each positive example paired with 4 negative examples, and output underwent L2 normalization to enhance representation quality. To accelerate the training process, mixed precision training (FP16), gradient accumulation (step = 4), and distributed training (8 GPUs) techniques were applied. Training experiments were conducted on a node configured with eight 32GB V100 GPUs, supplemented by a high-performance computing environment with 32-core CPU and 384GB memory.

In practical implementation, this research trained embedding subspaces at four nested scales including 128, 256, 512, and 1024 dimensions, forming a hierarchical representation structure. In system deployment, the maximum scale 1024-dimensional embedding is typically used for retrieval to obtain highest precision; however, in resource-constrained scenarios, the system can seamlessly switch to smaller dimensions, such as 512 or 256 dimensions, significantly reducing computational and storage costs while maintaining comparable retrieval quality.

5. Training Infrastructure

The training process was conducted on a cluster of 14 machines with 8 NVIDIA V100 GPUs each (112 GPUs total, NVIDIA, Santa Clara, CA, USA), adopting a distributed data parallel strategy. Efficient inter-node communication was achieved through InfiniBand network. Training used MS-Swift as the core framework for distributed training. The embedding model fine-tuning was conducted on a single node with 8 V100 GPUs (32 GB each), supplemented by a 32-core CPU and 384 GB memory computing environment.

6. Data Extraction Methodology

Our data source prioritization follows: HTML format preferred, PDF as secondary source. Most academic publications provide HTML versions with cleaner structure, which are preferentially used for text extraction.

For PDF processing, the PyMuPDF (fitz) library is employed for extracting text from native digital PDFs. Our pipeline processes exclusively text-based (digitally born) PDFs. Scanned documents that cannot be parsed by PyMuPDF are identified and excluded from the corpus, as the academic publications in our target domain are predominantly available in native digital format. Chemical formulas are preserved in standard notation (e.g., CH₄, H₂O, Pt/Al₂O₃), mathematical expressions are retained in LaTeX format where possible, and subscripts/superscripts are converted to standard text representation.

Regarding non-textual content, tables are extracted and formatted as HTML structure, preserving row/column relationships. For figures and schematics, only captions are utilized in the current version. We acknowledge that visual information in figures is not captured, which represents a direction for future multi-modal extension.

Quality control measures include DOI-based deduplication, rule-based cleaning (removing headers, footers, and reference sections), and LLM-assisted quality filtering for coherence and completeness.

7. Extended Retrieval Performance Evaluation

Table S2 below presents the extended retrieval performance metrics. The fine-tuned embedding model (WithMrl) achieves an average Recall@10 of 66.65% on propane dehydrogenation validation datasets (English: 77.31%, Chinese: 55.98%), compared to the bge-m3 baseline's 63.95% (English: 76.05%, Chinese: 51.84%), representing a 2.70 percentage point improvement. On general document QA tasks, the fine-tuned model achieves Recall@3 of 83.49% and Recall@10 of 94.70%, compared to the baseline's 81.93% and 94.08% respectively, demonstrating that domain-specific fine-tuning does not compromise general retrieval capabilities.

Table S2. Extended retrieval performance metrics.

Model	Dataset	Precision@1	Precision@3	Precision@5	Precision@10	Recall@1	Recall@3	Recall@5	Recall@10
Fine-tuned Model	PDH-zh	30.51%	14.88%	9.89%	5.60%	30.51%	44.63%	49.45%	55.98%
jina-v2-base-zh	PDH-zh	19.24%	10.06%	7.00%	4.15%	19.24%	30.19%	34.98%	41.53%
bge-m3	PDH-zh	26.89%	13.39%	9.06%	5.18%	26.89%	40.17%	45.31%	51.84%
bge-large-zh-v1.5	PDH-zh	13.60%	7.42%	5.21%	3.15%	13.60%	22.25%	26.03%	31.52%
bge-large-zh	PDH-zh	11.60%	6.54%	4.71%	2.89%	11.60%	19.61%	23.53%	28.91%
bge-base-zh-v1.5	PDH-zh	12.52%	6.71%	4.76%	2.85%	12.52%	20.14%	23.81%	28.45%
bge-base-zh	PDH-zh	10.09%	5.74%	4.06%	2.52%	10.09%	17.21%	20.31%	25.22%
Fine-tuned Model	PDH-en	51.30%	22.17%	14.23%	7.73%	51.30%	66.51%	71.13%	77.31%
jina-v2-base-zh	PDH-en	37.69%	17.40%	11.65%	6.52%	37.69%	52.20%	58.23%	65.22%
bge-m3	PDH-en	49.55%	21.43%	13.95%	7.60%	49.55%	64.28%	69.76%	76.05%
bge-large-zh-v1.5	PDH-en	23.34%	11.24%	7.65%	4.48%	23.34%	33.71%	38.27%	44.77%
bge-large-zh	PDH-en	23.56%	11.55%	7.89%	4.65%	23.56%	34.64%	39.46%	46.54%
bge-base-zh-v1.5	PDH-en	24.13%	11.61%	7.89%	4.56%	24.13%	34.84%	39.43%	45.59%
bge-base-zh	PDH-en	21.91%	10.51%	7.17%	4.21%	21.91%	31.52%	35.87%	42.08%
Fine-tuned Model	General	68.54%	27.83%	17.76%	9.47%	68.54%	83.49%	88.79%	94.70%
jina-v2-base-zh	General	54.52%	23.88%	15.39%	8.32%	54.52%	71.65%	76.95%	83.18%
bge-m3	General	68.22%	27.31%	17.45%	9.41%	68.22%	81.93%	87.23%	94.08%
bge-large-zh-v1.5	General	59.19%	25.75%	16.45%	8.79%	59.19%	77.26%	82.24%	87.85%
bge-large-zh	General	52.02%	22.12%	14.89%	8.35%	52.02%	66.36%	74.45%	83.49%
bge-base-zh-v1.5	General	60.44%	25.13%	16.20%	8.82%	60.44%	75.39%	81.00%	88.16%
bge-base-zh	General	49.22%	22.12%	14.27%	7.69%	49.22%	66.36%	71.34%	76.95%
Fine-tuned Model	Computational-zh	26.61%	12.23%	8.05%	4.48%	26.61%	36.69%	40.25%	44.81%
jina-v2-base-zh	Computational-zh	16.37%	8.04%	5.49%	3.19%	16.37%	24.13%	27.45%	31.87%
bge-m3	Computational-zh	23.26%	10.94%	7.31%	4.12%	23.26%	32.83%	36.57%	41.22%
bge-large-zh-v1.5	Computational-zh	11.05%	5.53%	3.81%	2.24%	11.05%	16.60%	19.05%	22.41%
bge-large-zh	Computational-zh	9.06%	4.58%	3.19%	1.91%	9.06%	13.73%	15.94%	19.14%
bge-base-zh-v1.5	Computational-zh	10.47%	5.21%	3.58%	2.10%	10.47%	15.64%	17.88%	20.99%
bge-base-zh	Computational-zh	7.70%	3.90%	2.70%	1.61%	7.70%	11.71%	13.50%	16.13%
Fine-tuned Model	Computational-en	50.15%	21.27%	13.60%	7.27%	50.15%	63.80%	67.99%	72.70%
jina-v2-base-zh	Computational-en	36.69%	16.50%	10.82%	5.96%	36.69%	49.51%	54.10%	59.64%
bge-m3	Computational-en	49.07%	21.02%	13.49%	7.26%	49.07%	63.07%	67.43%	72.57%
bge-large-zh-v1.5	Computational-en	19.51%	9.07%	6.11%	3.50%	19.51%	27.22%	30.55%	35.03%
bge-large-zh	Computational-en	20.00%	9.43%	6.38%	3.67%	20.00%	28.30%	31.92%	36.72%
bge-base-zh-v1.5	Computational-en	20.76%	9.50%	6.37%	3.64%	20.76%	28.50%	31.86%	36.35%
bge-base-zh	Computational-en	17.62%	8.28%	5.59%	3.23%	17.62%	24.84%	27.97%	32.28%
Fine-tuned Model	Thermal Catalysis-zh	23.92%	11.42%	7.64%	4.34%	23.92%	34.27%	38.21%	43.45%
jina-v2-base-zh	Thermal Catalysis-zh	13.16%	6.77%	4.73%	2.82%	13.16%	20.31%	23.66%	28.24%
bge-m3	Thermal Catalysis-zh	20.24%	9.88%	6.72%	3.88%	20.24%	29.63%	33.59%	38.82%
bge-large-zh-v1.5	Thermal Catalysis-zh	9.28%	4.79%	3.35%	2.02%	9.28%	14.38%	16.77%	20.24%
bge-large-zh	Thermal Catalysis-zh	7.16%	3.77%	2.70%	1.65%	7.16%	11.32%	13.48%	16.51%
bge-base-zh-v1.5	Thermal Catalysis-zh	8.58%	4.39%	3.07%	1.85%	8.58%	13.16%	15.35%	18.52%
bge-base-zh	Thermal Catalysis-zh	6.29%	3.33%	2.35%	1.44%	6.29%	9.98%	11.77%	14.44%
Fine-tuned Model	Thermal Catalysis-en	44.33%	19.32%	12.53%	6.83%	44.33%	57.97%	62.64%	68.28%
jina-v2-base-zh	Thermal Catalysis-en	31.05%	14.37%	9.58%	5.40%	31.05%	43.11%	47.89%	53.99%
bge-m3	Thermal Catalysis-en	42.96%	18.90%	12.33%	6.75%	42.96%	56.70%	61.67%	67.52%
bge-large-zh-v1.5	Thermal Catalysis-en	16.63%	7.92%	5.40%	3.15%	16.63%	23.77%	27.01%	31.48%
bge-large-zh	Thermal Catalysis-en	16.39%	7.93%	5.45%	3.20%	16.39%	23.79%	27.23%	31.99%
bge-base-zh-v1.5	Thermal Catalysis-en	17.96%	8.41%	5.71%	3.31%	17.96%	25.23%	28.56%	33.15%
bge-base-zh	Thermal Catalysis-en	14.96%	7.19%	4.93%	2.90%	14.96%	21.58%	24.66%	28.96%

The lower Chinese retrieval performance (Recall@10 of 55.98% vs. 77.31% for English) is primarily attributed to the composition of the training corpus. The industrial catalysis literature is predominantly published in English, resulting in significantly fewer Chinese-language documents in both the knowledge base and the retrieval training set. This inherent corpus imbalance directly impacts Chinese retrieval metrics. Furthermore, cross-lingual embedding alignment for technical domains remains an open challenge. Aligning semantic representations across languages is particularly difficult for specialized terminology in industrial catalysis, where Chinese technical terms often have domain-specific conventions that do not map directly to their English counterparts. Importantly, in the deployed system, this lower Chinese embedding recall does not proportionally impact end-user experience. The query processing pipeline includes automatic query translation, enabling Chinese queries to be simultaneously searched against the predominantly English knowledge base in English. Combined with the dual-path retrieval mechanism (vector search + BM25) and result fusion, the effective retrieval coverage for Chinese-language queries is substantially higher than the single-path Chinese Recall@10 metric suggests.

8. Iterative Retrieval Convergence Behavior

During system testing, we observed that the first round of RAG retrieval resolves the majority of inconsistency cases between the model's initial answer and the retrieval-augmented answer. In our test set, after the first RAG iteration, the enhanced answer and the newly retrieved supporting literature achieved consistency in all observed cases—no instance required the second iteration to further revise the answer. The second iteration, when triggered, consistently served to supplement additional corroborating literature rather than to correct the first-round RAG answer. While this observation suggests strong convergence properties of the iterative mechanism, we note that this finding is based on our current test set and may not generalize to all possible query types. A more rigorous statistical analysis of convergence rates across diverse query categories is planned for future work.

9. Domain Fine-tuning Effect on Question-Answering Quality

Figure S4 presents question-answering responses from the Yi-1.5-6B-Chat model before domain fine-tuning. For the identity query ('Who are you?'), the model identifies itself as Yi developed by 01.AI. For the domain-specific query ('What are PDH alloy catalysts?'), the model produces a fundamentally incorrect response, misinterpreting PDH as 'Partial Differential Hydrogen' and describing hydrogen production catalysts instead of propane dehydrogenation catalysts. This demonstrates the limitations of general-purpose models in specialized domain knowledge.

Figure S5 presents question-answering responses from the PeiYang Micro-Emergence model after domain fine-tuning. For the identity query, the model correctly identifies itself as developed by the Energy & Catalysis Adventure Team at Tianjin University. More importantly, for the same domain-specific query ('What are PDH alloy catalysts?'), the fine-tuned model correctly identifies PDH as propane dehydrogenation and provides an accurate, detailed response covering specific alloy catalyst systems (PtSn, PtIn, PtGa, PtZn) with appropriate technical context. This comparison demonstrates that domain fine-tuning fundamentally improves the model's understanding of specialized catalysis terminology and concepts.

Intelligent Question-Answering System for Industrial Catalysis
Powered by Yi-1.5-6B-Chat model

USER QUESTION
Who are you, and who developed you?

AI RESPONSE
I am Yi, a large-scale language model independently developed by 01.AI. I can answer questions, provide information, discuss topics, write articles, and more, aiming to help people obtain information more conveniently and effectively, and to engage in logical and emotionally rich conversations with people.

USER QUESTION
What are PDH alloy catalysts?

AI RESPONSE
PDH alloy catalyst (PDH stands for Partial Differential Hydrogen) is a high-efficiency catalyst used for hydrogen production. It is commonly used in steam methane reforming processes to convert methane (CH₄) and hydrogen (H₂) into carbon monoxide (CO), carbon dioxide (CO₂), water, and hydrogen. The design of this catalyst aims to enhance the activity and selectivity of the reaction, thereby increasing hydrogen yield while reducing the generation of byproducts.
PDH alloy catalysts are typically composed of metals or alloys with high specific surface area and excellent catalytic activity, such as cobalt, nickel, iron, copper, and aluminum. These metals or alloys exist in different forms within the catalyst, such as oxides, hydroxides, and sulfides, to provide optimal catalytic performance.
In the steam methane reforming process, the role of the catalyst is to lower the activation energy of the reaction, promote interactions between reactant molecules, and thus facilitate hydrogen production. PDH alloy catalysts can also regulate the reaction rate, ensuring that the reaction proceeds under the required temperature and pressure conditions.
Due to the role of PDH alloy catalysts in improving hydrogen production efficiency, they have been widely applied in the energy and chemical industries. As the demand for clean energy and sustainable development increases, the research and development of PDH alloy catalysts continues to advance, aiming to improve catalyst stability and durability while reducing environmental impact.

Start conversing with Yi-1.5-6B-Chat model, use Shift+Enter for line breaks

Figure S4. Question-answering responses from Yi-1.5-6B-Chat model before domain fine-tuning, showing incorrect interpretation of domain-specific terminology (PDH misidentified as ‘Partial Differential Hydrogen’).

Intelligent Question-Answering System for Industrial Catalysis
Powered by PeiYang Micro-Emergence model

USER QUESTION
Who are you, and who developed you?

AI RESPONSE
I am an artificial intelligence language model developed by the Energy & Catalysis Adventure Team at Tianjin University, known as PeiYang Micro-Emergence.

USER QUESTION
What are PDH alloy catalysts?

AI RESPONSE
PDH alloy catalysts are catalysts used for propane dehydrogenation (PDH) reactions, which represent a promising method for propylene production. PDH is an endothermic reaction that requires high temperatures to achieve high propylene yields.
PDH alloy catalysts are typically composed of two or more metals that can exist in alloy form, creating alloy phases with specific structural and electronic properties. These alloy phases can enhance the activity and selectivity of the catalyst, improving its performance in PDH reactions.
Some examples of PDH alloy catalysts include:
1. PtSn alloy catalysts: PtSn alloy catalysts are among the most extensively studied catalysts for PDH reactions. PtSn alloy catalysts are typically composed of Pt and Sn, forming alloy phases with specific structural and electronic properties. These alloy phases can enhance the activity and selectivity of the catalyst, improving its performance in PDH reactions.
2. PtIn alloy catalysts: PtIn alloy catalysts are another promising catalyst for PDH reactions. PtIn alloy catalysts are typically composed of Pt and In, forming alloy phases with specific structural and electronic properties. These alloy phases can enhance the activity and selectivity of the catalyst, improving its performance in PDH reactions.
3. PtGa alloy catalysts: PtGa alloy catalysts are another promising catalyst for PDH reactions. PtGa alloy catalysts are typically composed of Pt and Ga, forming alloy phases with specific structural and electronic properties. These alloy phases can enhance the activity and selectivity of the catalyst, improving its performance in PDH reactions.
4. PtZn alloy catalysts: PtZn alloy catalysts are another promising catalyst for PDH reactions. PtZn alloy catalysts are typically composed of Pt and Zn, forming alloy phases with specific structural and electronic properties. These alloy phases can enhance the activity and selectivity of the catalyst, improving its performance in PDH reactions.
In summary, PDH alloy catalysts are promising catalysts for PDH reactions that can enhance catalyst activity and selectivity, improving performance in PDH reactions.

Start conversing with PeiYang Micro-Emergence model, use Shift+Enter for line breaks

Figure S5. Question-answering responses from PeiYang Micro-Emergence model after domain fine-tuning, demonstrating accurate domain knowledge with correct interpretation of PDH (propane dehydrogenation) and detailed catalyst information.

10. Practical Generation Examples of the Overall Workflow

To illustrate the complete retrieval-augmented generation workflow described in the main text, Figure S6 presents practical example responses from the deployed system, demonstrating how the system integrates domain-specific model capabilities with retrieval augmentation to deliver accurate and literature-grounded professional knowledge services.

Intelligent Question-Answering System for Industrial Catalysis
Powered by PnYang Micro-Emergence model

USER QUESTION
 What're the pros and cons of these Ni-based propene dehydrogenation catalysts anyway?

AI RESPONSE
 Nickel-based catalysts exhibit high activity in propene dehydrogenation reactions, but suffer from several drawbacks. Firstly, nickel-based catalysts are prone to deactivation due to carbon deposition, leading to short catalyst lifetime. Secondly, nickel-based catalysts demonstrate relatively low propylene selectivity, as they readily promote deep dehydrogenation reactions that generate coke. Furthermore, the stability of nickel-based catalysts is limited, as they are susceptible to sintering and agglomeration. Therefore, it is necessary to develop nickel-based catalysts with enhanced activity and selectivity to improve the efficiency of propene dehydrogenation reactions.

Thinking Process

Rewrite Input Query

What are the advantages and disadvantages of nickel-based propene dehydrogenation catalysts?

Generate Initial Results

Generating initial results...

Translate Input Content

What are the advantages and disadvantages of nickel-based propene dehydrogenation catalysts?

Generate Translation Results

Nickel-based catalysts are commonly used in propene dehydrogenation (PDH) processes due to their high activity and selectivity. However, there are also some disadvantages associated with these catalysts.

Advantages of nickel-based catalysts:

- High activity:** Nickel-based catalysts have a high activity for the dehydrogenation of propene to propylene. This allows them to achieve high propylene yields at relatively low temperatures.
- High selectivity:** Nickel-based catalysts are also known for their high selectivity towards propylene production. This means that they are less likely to produce byproducts such as methane, ethane, and ethylene.
- Cost-effectiveness:** Nickel-based catalysts are relatively inexpensive compared to other catalysts used in PDH processes. This makes them an attractive option for industrial applications.
- Versatility:** Nickel-based catalysts can be used in a variety of PDH processes, including fixed-bed reactors, fluidized-bed reactors, and membrane reactors. This versatility allows them to be used in different industrial settings.

Disadvantages of nickel-based catalysts:

- Limited stability:** Nickel-based catalysts can suffer from deactivation over time due to the formation of carbon deposits on the catalyst surface. This deactivation can be accelerated by factors such as high temperatures, low H₂/C₃H₆ ratios, and the presence of impurities in the feed.
- High cost:** While nickel-based catalysts are relatively inexpensive compared to other catalysts used in PDH processes, they can still be a significant cost for industrial applications.
- Environmental impact:** Nickel-based catalysts can release toxic compounds such as nickel and carbon dioxide into the atmosphere during the PDH process. This can have negative environmental impacts.
- Limited catalyst performance:** Nickel-based catalysts can have limited catalytic performance compared to other catalysts used in PDH processes. This can be due to factors such as the presence of impurities in the feed, the presence of carbon deposits on the catalyst surface, or the presence of other metals that can interact with the nickel active phase.

In summary, nickel-based catalysts are widely used in PDH processes due to their high activity and selectivity. However, they also have some disadvantages, including limited stability, high cost, environmental impact, and limited catalytic performance. These disadvantages must be carefully considered when selecting a PDH catalyst.

Related Literature Excerpts

DOI [1] 10.1021/acs.jpcc.8b03939
Journal: The Journal of Physical Chemistry C Year: 2018 Impact Factor: 3.1000 First Author: Tanshuo Sarker Last Author: Anshul Jadhav

Among various choices of reactive catalysts, Ni-based catalyst has been widely used in various catalytic applications, such as steam reforming processes, (43,44) water-gas shift reaction, (45,46) and biomass gasification. (47) Ni-based catalyst has also been reported as a good catalyst for hydrogenation (48) and dehydrogenation of small alkanes. (49-51) Recently, this catalyst has been reported as a candidate for the PDH process due to its high reactivity for hydrocarbon reaction, co-feedstocks, and lower cost compared to Pt-based catalysts. (26,43-52) Yun et al. (20) experimentally studied PDH reaction on Ni supported by SiO₂ (Ni/SiO₂). They found that the reaction rate of propene conversion is very high (about 1 × 10⁴ molecules/s at 523 K). The high propene conversion refers to high reactivity of Ni/SiO₂. The reactivity of Ni-based catalysts for methanol conversion is higher than that of Pt-based catalysts. (53-55) Although the reactivity of propene conversion on Ni/SiO₂ is very high, C₂H₄ was detected as a main product from the C-C bond cracking. This predominant side reaction results in low selectivity of PDH in this catalyst. (20) Moreover, the deactivation of Ni/SiO₂ from coke formation induced by deep dehydrogenation resulting in low propylene production was also reported. (20) The low selectivity of PDH on Ni was also reported in other experimental studies. For instance, Mikkola (56) proposed that the selectivity of propylene production might be interrupted by C-C bond cracking of propene during the PDH process due to the high reactivity of Ni toward the cracking reaction of alkane. Moreover, Ni-based catalysts have been reported as highly reactive catalysts for deep dehydrogenation leading to the creation of coking on Ni surface. (57,58) From the two possible side reactions of C-C bond cracking and deep dehydrogenation, the PDH mechanisms on Ni surface are still controversial at present. Understanding these reactions is necessary to develop better Ni-based catalysts for PDH.

DOI [2] 10.1016/j.jpc.2023.03.007
Journal: Journal of Pure and Applied Chemistry Year: 2023 Impact Factor: 5.2092 First Author: Lijun Hou Last Author: Aijun Duan

As an important high value-added chemical raw material, propylene plays an indispensable role in the petrochemical industry.¹ Traditionally, propylene is produced by catalytic cracking of petroleum fractions and by products (e.g., naphtha) or steam cracking of light diesel oil.^{2,3} However, the gap between supply and demand of propylene has gradually increased with the rapid consumption of fossil energy resources, and propylene prepared by conventional processes cannot satisfy the demand.⁴ Accordingly, the development of economical and highly efficient propylene production technologies is of critical importance. Since the widespread exploitation of shale gas around the world has brought more plentiful and cheaper propane, which makes propene dehydrogenation (PDH) to propylene become a more attractive alternative.⁵⁻⁶ Notably, compared with conventional processes, PDH technology has manifested advantages of low investment cost and high propylene yield.^{8,9} Currently, Pt- and CrOx-based catalysts are widely used for the catalytic dehydrogenation of propane.^{10, 11, 12, 13, 14} Especially catalysts such as Pt-Sn^{15,16} and CrAl₂O₃^{17,18} have been commercialized. Although such catalysts have shown high activity and selectivity, they suffer some drawbacks. Expensive platinum-based catalysts are easily deactivated due to carbon deposition, and toxic chromium-based catalysts cause serious environmental pollution. Therefore, it is particularly essential to develop and prepare green and inexpensive catalysts with high catalytic performance.

DOI [3] 10.3390/catal10091042
Journal: Catalysts Year: 2020 Impact Factor: 3.0001 First Author: Guangjun Wang Last Author: Dehua Hou

Propylene, as one of the most important types of compounds in the modern petrochemical industry, is mainly used to produce polypropylene, acrylonitrile, propylene oxide, and other important intermediates [1,2,3,4,5,6,7]. Propylene is traditionally supplied from the steam cracking and the catalytic cracking process of naphtha or other oil byproducts [8]. In recent years, the domestic and international market demand for propylene is dramatically growing with the continuous exploitation of propylene downstream products. Therefore, some on-propene propylene production technologies (e.g., methanol-to-olefins, methanol-to-propylene, propene dehydrogenation, etc.) are widely developed to satisfy the increasing demand for propylene. Among these technologies, the propene dehydrogenation (PDH), regarded as a potential and high value-added propylene production route, has gained much attention [9,10]. This process is an endothermic and thermodynamically equilibrium-limited reaction [9,11], which needs the operating conditions of high-temperature and low-pressure for realizing the high propylene yield. The high reaction temperature of 550–700 °C could lead to fast catalyst deactivation as well as unwanted side reactions, such as isomerization, hydrogenolysis, and dehydrocyclization [12,13]. Presently, the platinum-based catalyst (COP Oleflex process) and chromium oxide-based catalysts (Lumina Catalin process) were mainly employed in the commercial PDH processes [14]. However, the preparation and application of chromium oxide-based catalysts would be limited due to the environmental pollution issues.

DOI [4] 10.1016/j.apcata.2023.119167
Journal: Applied Catalysis A: General Year: 2023 Impact Factor: 4.7003 First Author: Huzail Kazemi Last Author: Ali Abbasi

Propylene is one of the most prominent gases due to some valuable products and derivatives such as polymers, solvents, dyes, etc., which makes it one of the most important building blocks in the chemical industry. Due to the limitations of steam cracking and fluid catalytic cracking processes in terms of low selectivity, energy consumption, and significant CO₂ emission, these processes cannot fulfill the growing demand for propylene [1], [2], [3]. Compared with traditional methods, catalytic propene dehydrogenation (PDH) has grabbed the attention of researchers due to the high yield of propylene, safety, and low operational cost. Pt and CrOx-based catalysts are commercial catalysts for PDH processes in Oleflex and CATOFIN technologies, respectively, and can achieve high activity and selectivity to olefins. In commercial PDH technologies, CrOx-based catalysts have been widely used due to their low cost, good mechanical stability, and outstanding performance [4], [5], [6].

Recommended Literature Links

DOI [1] 10.1016/j.cata.2020.104006
Journal: Catalysis Communications Year: 2020 Impact Factor: 3.4000 First Author: Fatima Omidihejazi Last Author: Esmail V. Kouhdad

DOI [2] 10.1016/j.apcatb.2020.121028
Journal: Catalysis Today Year: 2021 Impact Factor: 5.2092 First Author: Tejyinh Takher Last Author: Shokan M. Saffarzadeh

DOI [3] 10.1039/d0cc01017
Journal: Russian Chemistry & Engineering Year: 2020 Impact Factor: 3.4000 First Author: Shao Liu Last Author: Guohua Liu

DOI [4] 10.3390/catal10091042
Journal: Catalysts Year: 2020 Impact Factor: 3.0001 First Author: Guangjun Wang Last Author: Dehua Hou

DOI [5] 10.1021/acs.jpcc.8b03939
Journal: The Journal of Physical Chemistry C Year: 2018 Impact Factor: 3.1000 First Author: Tanshuo Sarker Last Author: Anshul Jadhav

DOI [6] 10.1016/j.cata.2020.104006
Journal: Catalysis Communications Year: 2020 Impact Factor: 3.4000 First Author: Fatima Omidihejazi Last Author: Esmail V. Kouhdad

DOI [7] 10.1016/j.apcata.2023.119167
Journal: Applied Catalysis A: General Year: 2023 Impact Factor: 4.7003 First Author: Huzail Kazemi Last Author: Ali Abbasi

DOI [8] 10.1016/j.cata.2021.101007
Journal: Catalysis Communications Year: 2021 Impact Factor: 3.4000 First Author: Z. Skordi Last Author: A.A. Lenzouari

DOI [9] 10.1016/j.jpc.2023.03.007
Journal: Journal of Pure and Applied Chemistry Year: 2023 Impact Factor: 5.2092 First Author: Lijun Hou Last Author: Aijun Duan

Start conversing with PnYang Micro-Emergence model, use Shift+Enter for line breaks

Figure S6. Example responses from the domain-specific retrieval-augmented generation system.